DOCUMENT RESUME

ED 423 281 TM 029 086

AUTHOR	Green, Donald Ross
TITLE	Why Is It So Hard To Agree on Professional Testing Standards? A Test-Publishing Perspective.
PUB DATE	1998-04-16
NOTE	9p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Opinion Papers (120) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Definitions; *Educational Testing; Elementary Secondary Education; Limited English Speaking; School Districts; Standards; State Legislation; *Test Construction; Test Validity; *Testing Problems
IDENTIFIERS	Opportunity to Learn; *Standards for Educational and Psychological Tests; *Test Publishers

ABSTRACT

As the new version of the "Standards for Educational and Psychological Testing" is being developed, it is apparent that putting together a set of standards for test publishing involves many difficulties. Although the basic intent of almost all parties involved is similar, there are many potential areas of disagreement among parties to the standards, which include test publishers, test sponsors, test users, and test takers. Areas of general agreement include the nature of validity and evidence for it, the importance of reliability, and general considerations of test construction. However, that areas of agreement about validity exist does not mean that agreement on specific wording is easily obtained. Test developers and school systems know that the uses of the test cannot be controlled by the developer. For this reason, asking the test developer for evidence of validity in specific situations is likely to be an area of disagreement. Trying to ensure that all necessary steps toward fairness and test bias have been taken is another area of potential disagreement about what the "Standards" should specify. Other issues come up in the section of the "Standards" related to educational testing that make it difficult for the parties to agree on what the "Standards" should say. One is the area of opportunity to learn. As it is being written, the standard for making decisions about student promotion or graduation requires that the test cover only what students have had the opportunity to learn. Determining what this is poses problems for all concerned. Other problems on which the "Standards" have little chance of gaining agreement are the testing of students of limited English proficiency and the tendency of state legislatures to pass testing requirements that cannot be met for one technical reason or another. It is difficult to agree on professional testing standards because the various groups have legitimately different interests and because the "Standards" are easy to misunderstand. Another reason is that the wording chosen can leave publishers, developers, sponsors, and users open to public and even legal attacks by those who dislike the outcomes or have political axes to grind. (Contains four references.) (SLD)

ED 423 281

Why Is It So Hard To Agree On Professional Testing Standards?

A Test-Publishing Perspective

Donald Ross Green

CTB/McGraw-Hill

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Donald Green

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the 1998 Annual Meeting of the American Educational Research Association

April 16, 1998
San Diego

TM029086

Introduction

It is certainly true that the task of putting together a set of standards for test publishing involves many difficulties. However, it is also important to recognize that there is widespread agreement about a large proportion of the concepts found in the various versions of the *Standards for Educational and Psychological Testing* (hereafter the *Standards*) of the past and in the new set of standards now nearing completion. The process of actually writing a set of standards that will be cited as gospel by one's professional colleagues, by courts of law, and by all sorts of people who probably do not really understand them is long and painful. Each word needs to be weighed to avoid ill-considered standards and to gain acceptance by as many relevant parties as possible. Having been a reviewer and critic of the 1975 and 1985 versions and of what I hope will be the 1998 version of the *Standards*, as well as having participated at length in the discussions of the *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 1988), I am well aware of the difficulty of obtaining assent (consensus is too strong a word), and I have great respect for the dedication and wisdom of the members of the current committee and of its predecessors.

The 1985 *Standards* (AERA/APA/NCME, 1985) assert that the interested parties are the test publishers (yes, we are interested in standards), the test sponsors (some are interested in standards), the test users (many, but far from all, are interested in standards) and the test takers (the overwhelming majority have never even heard of standards for test publishing).

Speaking as a representative of one of the most relevant parties, the publishers of educational tests, I would like to emphasize that I believe that the basic intent of almost all those involved is very similar. I also believe that many of the parties do not fully understand the concerns of others. It is to these potential sources of disagreement that I will direct most of my comments.

Of course, the members of the various groups do not all agree with other members of their own groups, but I believe that this is a much less serious problem because the political consequences are not as potent. Please note that the views expressed here have not been seen by, much less endorsed by, other publishers. Also the context of these remarks is largely educational measurement; the issues around tests used in clinical settings, licensure tests and employment tests have not been considered. Finally, not knowing exactly what the final words in the new version of the *Standards* will say about any particular point, these comments may end up making much of potential disagreements that will have become moot, while missing issues that will be contentious in the future.

The 1985 *Standards* mention yet another interested party, namely the academics who may review the tests, but do not note that almost all of the people on the committees, past and present, that write the *Standards* come from this group. Consulting, teaching, and conducting research about the issues dealt with in the *Standards* are the major activities of many, if not most, of the academics who specialize in educational and psychological measurement. Many of them also become involved in test development activities. They are not disinterested parties and obviously they do not all agree with each other (they wouldn't be academics if they did). Nevertheless, the resolutions to disagreements do not generally harm them even when the decisions go against them. That is not necessarily the case for the other groups.

Areas of Agreement

Before discussing the potential disagreements and misunderstandings that arise about the *Standards*, it may be well to indicate briefly how broad the areas of agreement are. I believe there is substantial agreement about most of the basic concepts. What follows in the next three paragraphs is a set of rather sweeping generalizations. I know that many counterexamples of each of them can be found, but I do believe that the generalizations tend to be true.

First is the nature of validity and evidence for it. From where I sit, it appears that the arguments about this topic tend be centered within the academic community. The disagreements among the various parties listed above tend to be about the amount of evidence needed (although everyone agrees that more evidence is always desirable) and about who is responsible for obtaining the evidence, but not about the kinds of evidence or the relevance of that evidence.

The situation appears to be similar with respect to reliability. Everybody agrees that reliability is important and that more is better, but they do not agree about the amount of evidence that is needed. In spite of the fact that reliability has been at the center of psychometrics for decades, and in spite of the fact that everybody thinks they understand it, many people take (or prefer to take) a very simplistic view of reliability well short of what those writing the *Standards* have in mind. But the nature of reliability evidence does not appear to have been a serious bone of contention among the several parties.

Although there are vigorous arguments about which tests have had the best construction procedures, people only rarely point to the *Standards* in these discussions. Similarly, although the amount of heat that can be found in discussions of scales and equating is sometimes quite surprising, the *Standards* do not appear to play much of role in these arguments. The standards concerning test administration, scoring, and reporting, as well as those referring to test documentation, do not arouse much emotion either, even though there are small arguments about the wording of these standards.

Areas of Potential Disagreement

Does this mean that I think writing the first six basic chapters of the *Standards* has been easier than writing the more narrowly focused chapters that follow? Not at all. I merely mean to make the point that the problems I have chosen to discuss stem, at least partly, from the particular choice of words and the presentation, rather than from fundamental disagreements about what kinds of standards tests should meet. Other problems stem from essentially intractable policy issues that technical standards cannot logically deal with. For the first type of problem, I will use some examples from the validity standards. For the second type of problem, I will use some examples from the fairness standards and from their ramifications in the educational testing standards.

Validity

General agreement does not mean that agreement on the specific wording is easily obtained. For example, Standard 1.1 in the 1985 *Standards* rarely if ever draws argument because it is so very broad that only those not offering any evidence of validity at all are likely to be viewed as not having met it. But suppose it were to say that a test developer must specify precisely how the test scores are intended to be interpreted and used and for which population. This would suggest a clearer situation than exists in many areas. For example, achievement test results are used in school system assessments by many different kinds of people: teachers, principals, superintendents, technical advisory committees, the media, parents, and even students (sometimes). Each of these groups has different uses in mind (cf. Taleporos, in press). Both the test developer and the school system know that various other uses will be made and that the uses and interpretations cannot be fully controlled.

Uses and interpretations of test results are often made that neither the school system nor the developer can know about ahead of time. Although a particular sort of interpretation may be suggested by the contract and the score reports designed with that use in mind, it can be assumed that many other uses and interpretations are going to arise. It is neither practical nor advisable for the developer to set forth clearly or disallow all these uses and interpretations. Not only are there too many of them, it could prove unwise to suggest, even negatively, some of these interpretations because someone could act on one or more of the inappropriate interpretations.

Neither the developer nor the sponsor can control what some users do with test results. It is often not possible to even know what is being done ahead of time, much less get the user to do validity studies. However, it seems likely that, if they were to be reproached for not doing such studies, many users would protest that they have neither the resources nor the expertise to do validity studies and that they did not (and perhaps do not) understand the need. In short, the wording about what happens when users take the bit in their teeth is almost certain to be deemed inappropriate by one group or another.

The need to have people assume responsibility for obtaining validity evidence is a basic idea almost all agree upon, but the writers of the *Standards* have put qualifying statements in the various introductions about using professional judgment in applying the *Standards* and about circumstances that one may not be able to control. These statements are welcome and can, in theory, prevent the use of the *Standards* to attack reasonable testing programs. However, claiming that professional judgment suggests a departure from some particular standard, or claiming that there were uncontrollable circumstances, is not an effective sort of defense when dealing with criticisms from outside groups. Such claims are just not heard and/or understood.

Another example is a newcomer to the domain of test validity, one that will, I am sure, appear in the forthcoming *Standards*, namely, concern with the consequential aspects of validity. As I noted in 1997 (Green, in press), the whole domain of the consequences of achievement testing programs is full of unverified and contradictory claims. Evidence is truly thin, even though faith in many of these claims is widespread. Publishers of nationally standardized achievement tests are distant both physically and in time from most of the possible evidence of the consequences of the uses and misuses of the tests. Ordinarily, publishers do not have access to the evidence and it does not even exist until long after the test has been planned and produced. Thus, any statement in the *Standards* that achievement test publishers, in contrast to test sponsors and especially test users, should be responsible for obtaining such evidence is problematic for publishers because it is not feasible to obtain it in any reasonable time frame.

Fairness

Some issues naturally lead to disagreements among groups, and the most prominent issues are those related to fairness and test bias. Test publishers and developers would like to believe that they have taken all the necessary steps as recommended by the *Standards* to eliminate bias and that any unfairness one might find is a consequence of misuse that was unforeseen and/or unpreventable by the developer. Test sponsors and users would like to believe that issues of fairness and test bias are the problem of the developer and publisher. Some groups that claim to represent test takers, such as Fairtest, believe that all those in the testing enterprise either unintentionally or intentionally ignore and refuse to recognize as a fact that most educational tests and ability tests are biased against females and/or ethnic groups such as African Americans. The members of the Association of Black Psychologists tend to believe that ignoring this alleged bias is intentional.

The 1985 *Standards* did not speak strongly on this topic; the two relevant standards, 3.5 and 3.10, are rated Conditional. Most achievement test publishers now take steps to meet these two standards, although to date, only CTB puts the follow-up data from the standardization in its Technical Reports. The revision of the *Standards* is going to have a whole chapter on this topic and will undoubtedly make stronger statements. However, if the revision does not specify that evidence of lack of test bias is necessary, the critics are going to be unhappy and ask for change. If the document does make such a demand, the

publishers and developers of achievement tests are going to say that that is not possible because nobody knows how to do that. Vetting a test for item bias ordinarily reduces possible test bias, but neither that step nor sensitivity reviews can guarantee that the test is completely unbiased. Recognition of this fact in the *Standards* would probably help bring reality to the discussion, but would not make anybody happy except those who would destroy the whole notion of testing and evaluation in favor of their own judgments and personal prejudices.

Educational Testing

In addition to the problems already discussed, a variety of other issues come up in the section on educational testing that make it difficult to get agreement on what the *Standards* should say.

One of them is an aspect of fairness peculiar to educational testing that is known as "opportunity to learn," which I believe will appear in the new version of the *Standards* pretty much as is, perhaps because it is derived from a court decision. (Obviously, judges and lawyers know better what should happen in schools than those who work with them or in them.) Standard 8.7 in the 1985 *Standards* states, "When a test is used to make decisions about student promotion or graduation, there should be evidence that the test covers only the specific or generalized knowledge, skills, and abilities that students have had the opportunity to learn." In my view, this standard creates a series of problems for all concerned. To list a few problems:

- It suggests that students in some schools can get promoted or graduate without being held to desired standards of knowledge and skill because they are not taught.
- It reduces pressure on schools and teachers to teach the prescribed curriculum because it is not tested.
- It requires test developers to find out what in fact is being taught in schools, which takes place behind classroom doors and is, therefore, not really knowable.
- It means that test developers cannot create tests that sample the content domain as specified by content experts.
- It means that the generalizability of the test scores may be severely limited, that is, their validity as measures of the full content domain is diminished.

Nevertheless, it is clearly unfair to students to hold them accountable for knowledge and skills they have had no chance to learn. These considerations make it well nigh impossible to write standards that people can agree upon.

There are a number of other problems in educational testing that I will mention only in passing upon which the *Standards* have little chance of getting much agreement. One is the use of tests presented in English to students who do not have full understanding of the language. The current controversy in California about the STAR testing program is an example. I believe that both the governor who is pushing this program and the district personnel who are protesting it can find language in the *Standards* to support their

positions. Similar disagreements are surely going to arise as efforts are made to include special education students in district and statewide testing programs.

Another source of problems is the tendency of state legislatures to pass legislation that requires testing program features such as development in one year, multiple and incompatible uses, linking to other tests such as NAEP, and linking to some test that permits international comparisons such as in TIMMS but that does not fit the curriculum specifications of the local test. In these instances, the resulting tests are much less likely to meet the requirements of the *Standards*, and fingers are then pointed at the developer or the sponsoring agency or both.

Conclusions

Why is it so hard to agree on professional testing standards? Not because there is no agreement about what good tests and good testing programs should be like. That agreement is widespread and broad.

It is difficult to agree on professional testing standards partly because the different groups in the overall enterprise have legitimately different interests; partly because the *Standards*, like the test results themselves, are so easy to misunderstand; and partly because the particular wording can sometimes leave the publishers and developers and/or the sponsors and users open to public and even legal attacks by those who dislike the outcomes or have political axes to grind.

References

AERA/APA/NCME. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Green, D.R. (in press). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice.*

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education.* Washington, DC: Author.

Taleporos, E. (in press). Consequential validity: A practitioner's perspective. *Educational Measurement: Issues and Practice.*